



Sveučilište u Zagrebu
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
Kemijски odsjek

Ana Mikelić

Studentica 1. godine Poslijediplomskog sveučilišnog studija KEMIJA,
smjer FIZIKALNA KEMIJA

MONTE CARLO PRISTUP MOLEKULARNOM DOKINGU

Kemijски seminar I

Izrađen prema:

H. Wang, H. Liu, L. Cai, C. Wang, and Q. Lv, "Using the multi-objective optimization replica exchange Monte Carlo enhanced sampling method for protein-small molecule docking," *BMC Bioinform*, 18(1), 1–21 (2017).

Mentor rada: prof. dr. sc. Tomica Hrenar

Zagreb, 2020.

Sadržaj

§ 1. UVOD.....	1
§ 2. MOLEKULARNI DOKING.....	2
2.1. Generiranje konfiguracija.....	2
2.2. Funkcije bodovanja.....	3
§ 3. MONTE CARLO METODE	4
3.1. Implementacija MC metode u RosettaLigand.....	5
§ 4. MONTE CARLO METODE UZ ZAMJENU REPLIKA.....	8
4.1. Implementacija REMC metode u RosettaLigand	9
§ 5. MONTE CARLO METODE UZ VIŠEKRITERIJSKU OPTIMIZACIJU ZAMJENE REPLIKA	12
5.1. Višekriterijska optimizacija	13
5.2. Implementacija MO-REMC u RosettaLigand	14
5.3. Implementacija hibridne MO-REMC u RosettaLigand.....	16
§ 6. USPOREDBA ODABRANIH MONTE CARLO METODA	18
§ 7. ZAKLJUČAK	23
§ 8. LITERATURNI IZVORI.....	XXIV

§ 1. UVOD

Interakcije bioloških makromolekula poput proteina, nukleinskih kiselina, lipida i ugljikohidrata međusobno kao i s drugim malim molekulama ključne su za pravilno provođenje biokemijskih procesa u tijelu. Osim fundamentalne uloge, poznavanje utjecaja vezanja jedne molekule (liganda) za drugu (receptor) važno je u istraživanju lijekova.¹

Vežanje liganda za receptor može se ispitati eksperimentalno, ali i računalnim metodama, između ostalog, molekularnim dokingom. Postupak započinje generiranjem svih mogućih konfiguracija sustava, za što još uvijek ne postoji idealna metoda. Novija istraživanja usmjerena su na povećanje brzine konvergencije poznatih metoda za generiranje konfiguracija. Kod brojnih metoda uzorkovanja (engl. *sampling methods*) pokazalo se da je cilj postignut korištenjem elitnih rješenja, važne komponente evolucijskih višekriterijskih optimizacijskih algoritama. Primjer jednog takvog poboljšanog algoritma uzorkovanja je Monte Carlo uz višekriterijsku optimizaciju zamjene replika (engl. *multi-objective optimization replica exchange Monte Carlo*, MO-REMC) razvijena od strane Wanga i sur.², koja u usporedbi s tradicionalnom Monte Carlo (MC) i MC metodom uz zamjenu replika (engl. *replica exchange Monte Carlo*, REMC) koristi informacije Pareto fronte višekriterijske optimizacije (engl. *multi-objective optimization Pareto front*) prilikom odabira replika za zamjenu. U ovom radu bit će detaljno objašnjenje spomenute metode te primjer njihove implementacije u komercijalni program za doking (RosettaLigand). Usporedba metoda bit će dana s obzirom na odabrane veličine iz RosettaLigand programa: ukupna vrijednost energijske funkcije (engl. *energy function total score, TScore*), razlika u veznoj energiji sučelja proteina (engl. *binding energy interface delta, IFDelta*) i RMSD liganda (engl. *ligand of RMSD, Lrmsd*).^{3,4}

§ 2. MOLEKULARNI DOKING

Molekularni doking obuhvaća računalne tehnike kojima se predviđaju načini vezanja liganda za receptor, najčešće supstrata u aktivno mjesto enzima. Primjenjuje se kada je poznata 3D-struktura molekule receptora, a želi se ispitati energija interakcije između te molekule i poznate ili potencijalne molekule liganda za različite međusobne rasporede i konformacije, odn. konfiguracije tih dviju molekula. Tradicionalno se provodi u dva koraka: generiranje svih mogućih konfiguracija te njihovo rangiranje na temelju pridružene funkcije bodovanja (engl. *scoring function*) i izdvajanje onih najrelevantnijih.¹

2.1. Generiranje konfiguracija

Generiranje mogućih konfiguracija može se provesti ručno putem kompjutorskog grafičkog sučelja u odgovarajućim programima ili primjenom automatskih algoritama. Problem u ovom koraku predstavljaju brojni stupnjevi slobode. Postoji šest stupnjeva slobode translacije i rotacije jedne molekule u usporedbi s drugom te konformacijski stupnjevi slobode obje molekule. Zadatak je stoga rješiv samo uz određene aproksimacije. Algoritmi za doking razlikuju se međusobno upravo prema stupnjevima slobode koje uzimaju u obzir.⁵

Najjednostavniji algoritmi zanemaruju konformacijske stupnjeve slobode i generiraju moguće konfiguracije na temelju rigidnih struktura promatranih molekula. Ne provodi se optimizacija geometrija već se promatra samo šest stupnjeva slobode translacije i rotacije.

Složeniji algoritmi pokušavaju uključiti što veći broj konformacijskih stupnjeva slobode. Većina njih promatra samo promjene konformacije liganda, dok se receptor tretira kao kruto tijelo. Koriste se svim uobičajenim metodama pretrage konformacijskog prostora.

Najrealnija slika dobila bi se omogućavanjem konformacijske fleksibilnosti i liganda i receptora. Za ovaj pristup nameće se primjena simulacija molekulske dinamike na intermolekulski kompleks liganda i receptora. Spomenuti računi kompjutacijski su zahtjevni i u praksi korisni samo za utočnjavanje struktura generiranih drugim metodama za doking.

2.2. Funkcije bodovanja

Iz velikog broja generiranih konfiguracija potrebno je izdvojiti one najznačajnije. Koliko je određena konfiguracija značajna može se kvantificirati pomoću funkcija bodovanja. Osim toga, one omogućavaju i usporedbu vezanja više različitih liganada za isti receptor.

Većina funkcija bodovanja pokušava aproksimirati Gibbsovu energiju vezanja liganda za receptor. Promatrani doprinosi Gibbsovoj energiji vezanja (ΔG_{bind}) navedeni su u jednadžbi:⁵

$$\Delta G_{\text{bind}} = \Delta G_{\text{solvent}} + \Delta G_{\text{conf}} + \Delta G_{\text{int}} + \Delta G_{\text{rot}} + \Delta G_{\text{t/r}} + \Delta G_{\text{vib}} \quad (1)$$

gdje $\Delta G_{\text{solvent}}$ uključuje doprinos solvatacijskih efekata zbog uravnoteživanja interakcija između otapala i liganda, proteina i intermolekulskog kompleksa, ΔG_{conf} dolazi od konformacijskih promjena u strukturi proteina i liganda, ΔG_{int} je Gibbsova energija specifičnih interakcija između proteina i liganda, ΔG_{rot} je smanjenje Gibbsove energije povezano sa smrzavanjem internih rotacija proteina i liganda, $\Delta G_{\text{t/r}}$ gubitak translacijske i rotacijske Gibbsove energije uzrokovano spajanjem dvaju tijela (liganda i receptora) u jedno tijelo (intermolekulski kompleks) i ΔG_{vib} doprinos Gibbsovoj energiji uzrokovan promjenama u načinima vibriranja.

§ 3. MONTE CARLO METODE

Metode generiranja početnih konfiguracija koje uzimaju u obzir fleksibilnost liganda mogu se podijeliti u 3 osnovne kategorije:⁶

1. Sistematske metode (engl. *systematic methods*): inkrementalno građenje (engl. *incremental construction*), konformacijska pretraga (engl. *conformational search*), baze podataka (engl. *databases*).
2. Nasumične ili stohastičke metode (engl. *random or stochastic methods*): Monte Carlo, genetski algoritmi (engl. *genetic algorithms*), tabu pretraživanje (engl. *tabu search*).
3. Metode simulacije (engl. *simulation methods*): molekulska dinamika (engl. *molecular dynamics*), energijska minimizacija (engl. *energy minimization*).

U skupinu nasumičnih ili stohastičkih metoda ubrajaju se algoritmi koji generiraju konfiguracije uvođenjem nasumičnih promjena kod liganda. Značajnost nove konfiguracije procjenjuje se na temelju unaprijed definirane funkcije vjerojatnosti.

Monte Carlo (MC) metode temelje se na Metropolis MC algoritmu, koji daje kriterij prihvaćanja nasumične promjene. U suštini se MC algoritmi sastoje od sljedećih koraka:^{6,7}

1. Generiranje početne konfiguracije nasumičnom rotacijom veza. U slučaju dokinga, za dobivanje početne konfiguracije dodatno se uključuje nasumična translacija i rotacija male molekule unutar aktivnog mjesta proteina.
2. Energijsko bodovanje početne konfiguracije.
3. Generiranje nove konfiguracije i bodovanje.
4. Prihvaćanje ili odbijanje nove konfiguracije na temelju Metropolis kriterija.
5. Ponavljanje prethodnih koraka dok se ne dobije željeni broj konfiguracija.

Metropolis kriterij nalaže da se novo rješenje automatski prihvaća ako ostvaruje bolje energijsko bodovanje od prethodnog rješenja. U suprotnom se računa funkcija vjerojatnosti nalik Boltzmannovoj raspodjeli:

$$P \sim \exp [-(E_1 - E_0)k_B T] \quad (2)$$

gdje su E_0 i E_1 energije liganda prije i nakon nasumične promjene, k_B Boltzmannova konstanta, a T termodinamička temperatura sustava. Pokaže li se nakon toga novo rješenje ipak značajnim,

ono se prihvaća i koristi kao početna konfiguracija za generiranje sljedeće; u suprotnom se konfiguracija odbacuje.^{8,9}

Vrijedi istaknuti da Metropolis algoritam generira Markovljev lanac stanja. Općenito, proces je Markovljev ako zadovoljava sljedeća dva uvjeta:⁵

1. Rezultat svakog novog pokušaja ovisi isključivo o prethodnom pokušaju.
2. Svaki pokušaj pripada konačnom skupu mogućih rezultata.

Razne varijante opisanog originalnog Metropolis algoritma implementirane su u brojne programe za doking. Primjerice u ranijim verzijama programa AutoDock,^{10,11} ICM,¹² QXP,¹³ MCDOCK,¹⁴ AutoDock Vina,¹⁵ i ROSETTALIGAND.³

3.1. Implementacija MC metode u RosettaLigand²

Integriranje MC simulacijom općenito omogućava procjenu $E[f(x)]$ uzimanjem uzoraka $\{X_t, t = 1, \dots, n\}$ iz prostora stanja Ω i zatim aproksimiranjem

$$E[f(x)] \approx \frac{1}{n} \sum_{t=1}^n f(X_t). \quad (3)$$

Vidljiva je zajednička značajka svih MC metoda, srednja vrijednost funkcije $f(X)$ procjenjuje se na temelju srednje vrijednosti uzorka. Kada su uzorci $\{X_t\}$ međudusono neovisni, zakon velikih brojeva osigurava valjanost primjene aproksimacije.

Algoritam 1 predstavlja pseudokod za RosettaLigand implementaciju MC Boltzmann metode uzorkovanja. Početna konformacija sustava (kompleks proteina i liganda) izmijeni se primjenom perturbacije *rotamer-TrialMover()* ili *packRotamersMover()*. Ovaj probni prijelaz *Mover()* iz zadnjeg prihvaćenog stanja (*old*) u perturbirano stanje (*new*) prihvaća se na temelju odgovarajuće vjerojatnosti prihvaćanja:

$$prob [old \rightarrow new] := e^{\min(40.0, \max(-40.0, boltz_factor))} \quad (4)$$

uz $boltz_factor = (last_accepted_score - score)/k_B T$, gdje je *last_accepted_score* energija zadnje prihvaćene strukture kompleksa, *score* energija perturbiranog sustava, T trenutna temperatura, a k_B Boltzmannovu konstantu, za koju se uzima da iznosi jedan. Probni prijelaz *Mover()* prihvaća se ili odbija na sljedeći način. Generira se nasumičan broj, označen s *mc_RG_uniform*, iz jednolike raspodjele na intervalu $[0,1]$. Pri čemu je vjerojatnost da je *mc_RG_uniform* $[0, 1]$ manji od $prob[old \rightarrow new]$ jednaka $prob[old \rightarrow new]$. Probni prijelaz *Mover()* prihvaća se ako je *mc_RG_uniform* $[0, 1] < prob[old \rightarrow new]$ ili $prob[old \rightarrow new] \geq 1$, a u suprotnom se odbacuje. Vjerojatnost prijelaza iz konformacije p u perturbiranu

konformaciju p' ovisi o razlici u $last_accepted_score - score$ između zadnjeg prihvaćene (*old*) konfiguracije i perturbirane (*new*) konfiguracije, koja je definirana tako da

$$P[p \rightarrow p'] := \begin{cases} 0, & \text{ako } prob[old \rightarrow new] \leq mc_RG_uniform[0,1], \\ 1, & \text{ako } prob[old \rightarrow new] > mc_RG_uniform[0,1], \\ 1, & \text{ako } prob[old \rightarrow new] \geq 1. \end{cases} \quad (5)$$

gdje je $prob[old \rightarrow new]$ vjerojatnost prihvaćanja prijelaza između konfiguracija p' i p . Ovo pravilo osigurava da je vjerojatnost prihvaćanja probnog prijelaza $Mover()$ iz zadnje prihvaćene konfiguracije u perturbiranu konfiguraciju zaista jednaka $prob[old \rightarrow new]$. Ako je nova konformacija odbačena, algoritam može zadržati dodanu kopiju prethodne prihvaćene strukture kao primjerak zadržan od strane sustava. **Slika 1** (gore lijevo) pokazuje da se u MC metodi zadnja pronađena struktura (crvena točka) prihvaća kao jedinstveno rješenje. Nakon većeg broja iteracija, može se dobiti pouzdana prosječna vrijednost energije promatranog kompleksa.

Algoritam 1: MCBOLTZMANN(p, T)

Input: p – trenutna struktura kompleksa, T – temperatura trenutnog sustava,

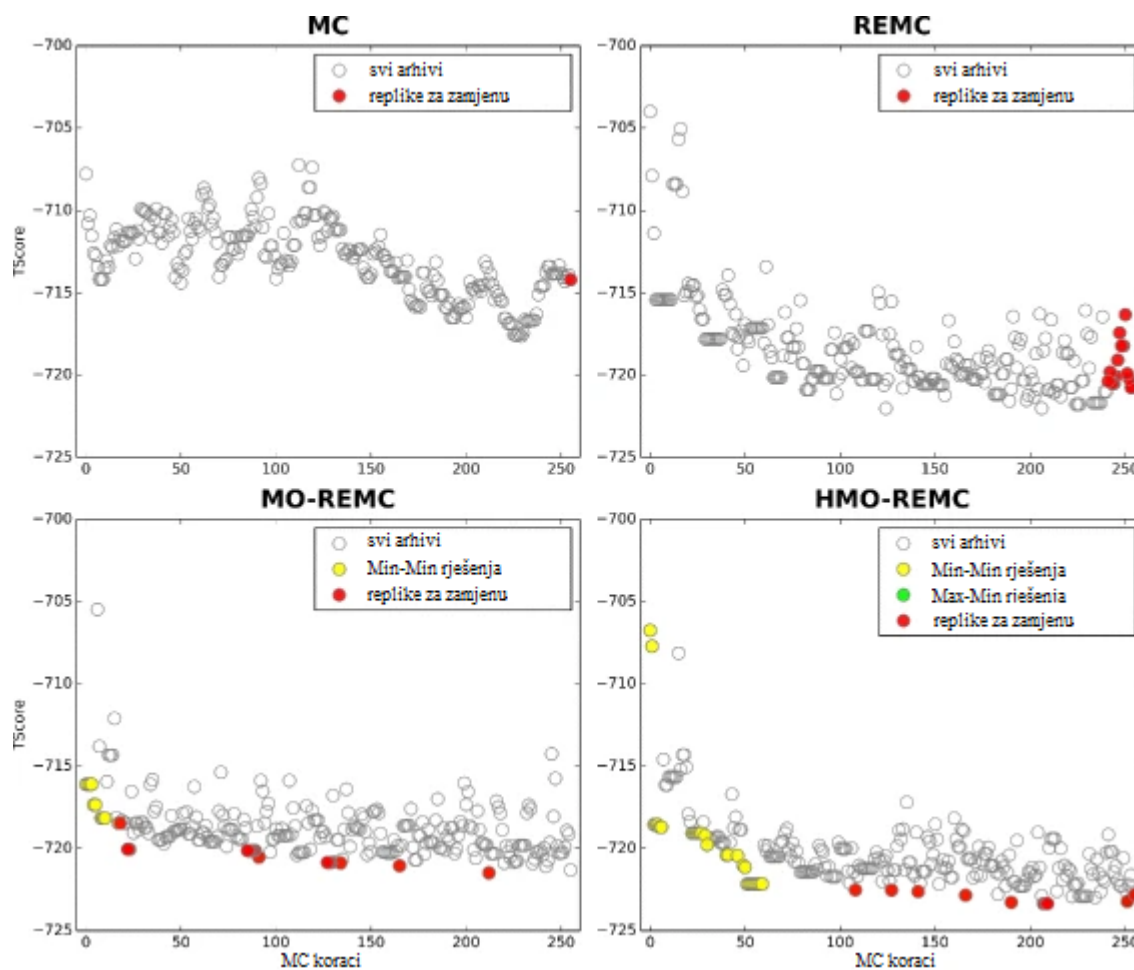
$E()$ – dodijeljena energijska funkcija

Output: $mc_accepted$ – točno ili netočno, dano prihvaćanje ili odbijanje trenutne strukture

```

1  $mc\_accepted \leftarrow 0$ 
2  $score \leftarrow E(p)$ 
3  $boltz\_factor \leftarrow (last\_accepted\_score - score) / T$ 
4  $prob \leftarrow e^{\min(40.0, \max(-40.0, boltz\_factor))}$ 
5 if  $prob < 1$  then
6    $mc\_RG\_uniform \leftarrow U(0,1)$ 
7   if  $mc\_RG\_uniform > prob$  then
8      $mc\_accepted \leftarrow 0$ 
9   else  $mc\_accepted \leftarrow 1$ 
10  else
11     $mc\_accepted \leftarrow 1$ 
12  end
13  if  $mc\_accepted$  then
14     $last\_accepted\_score \leftarrow score$ 
15  end
16  return  $mc\_accepted$ 

```

Slika 1. Replike odabrane u jednoj iteraciji pomoću MC, REMC, MO-REMC i HMO-REMC metode. (Preuzeto i prilagođeno iz ref. 2.)

§ 4. MONTE CARLO METODE UZ ZAMJENU REPLIKA

Poznato je da simulacije koje se provode pri nižim temperaturama imaju tendenciju ostati zarobljene u jednom od velikog broja lokalnih energijskih minimuma te je stoga pri niskim temperaturama teško dobiti pouzdanu kanonsku raspodjelu struktura tradicionalnim MC metodama. Jedan način kako riješiti ovaj problem višestrukih minimuma (engl. *multiple-minima problem*) je upravo provođenje simulacija na generaliziranom ansamblu (engl. *generalized ensemble*) koji se sastoji od niza stanja uteženih umjetnim utežnim faktorom ne-Boltzmannove vjerojatnosti tako da se može ostvariti nasumična šetnja prostorom potencijalne energije. Time je omogućeno prevladavanje svih energijskih barijera i uzorkovanje puno većeg konfiguracijskog prostora nego tradicionalnim metodama.¹⁶

U trenutnim protokolima, zamjena replika (engl. *replica exchange*) najkorištenija je metoda za poboljšanje uzorkovanja u računalnim simulacijama biomolekula. Metoda je razvijena kao nadogradnja simulacijskog kaljenja (engl. *simulation tempering*) uključivanjem generaliziranog ansambla te je poznata kao paralelno kaljenje (engl. *parallel tempering*) ili višestruki Markovljevi lanci (engl. *multiple Markov chains*). Ideja se bazira na hipotezi da je pri višim temperaturama potencijalna energija simuliranog sustava veća te da je sustavu lakše izaći iz lokalnog energijskog minimuma i istražiti globalni minimum. U MC metodi uz zamjenu replika (engl. *replica exchange Monte Carlo*, REMC) potencijalne energije paralelnih sustava (replika) simuliranih pri različitim temperaturama mogu se preklopiti. Rezultat toga je mogućnost zamjene konformacija niže energije iz visokotemperaturnih replika s konformacijama niskotemperaturnih replika. Time je omogućeno da se u normalnom temperaturnom rasponu mogu pretražiti strukture niže potencijalne energije pa je uzorkovanje efikasnije u usporedbi s tradicionalnom MC simulacijom.^{16,17}

Konkretno, M replika u kanonskim ansamblima neovisno i simultano simulira se na različitim temperaturama određeni broj MC koraka (ne postoji općenito pravilo koje bi određivalo optimalan broj koraka n između zamjena). Svaka temperaturna vrijednost jedinstvena je i svaka od M replika ima pridruženu temperaturu (T_1, T_2, \dots, T_M). Nakon n koraka, susjedne replike (npr. p_i i p_i') mogu zamijeniti konfiguracije na temelju određene vjerojatnosti

prijelaza. Ako dođe do zamjene, to podrazumijeva da replika p_i zamjenjuje konfiguraciju s replikom p_i' i s novom konfiguracijom nastavlja se simulacija na postavljenoj temperaturi.^{2,16}

Razne verzije REMC metode uzorkovanja koriste se u sklopu simulacija molekularnih sustava. Spomenute metode pretrage konformacijskog prostora kompjuterski su puno efikasnije od tradicionalnih MC metoda.²

4.1. Implementacija REMC metode u RosettaLigand²

REMC metoda implementirana korištenjem neovisnih Markovljevih lanaca p_i ($n \geq 0$) definirana je na jednakom prostoru stanja Ω kao MC (pogl. 3.1) i može se koristiti za paralelno ispitivanje većeg broja replika kako bi se istražile iste stacionarne normalne raspodjele $\rho_i(x)$ ($x \in \Omega$, $1 \leq i \leq N$) (zbog centralnog graničnog teorema) na različitim “temperaturama”. Zamjene se pokušavaju ostvariti između stanja p_i i p_i' dvaju „susjednih“ Markovljevih lanaca s indeksima i i j koji pripadaju različitim termodinamičkim stanjima, i tako mogu jednostavno biti identificirane konfiguracije koje prelaze lokalne energijske barijere.

Kao što je već spomenuto, REMC pretraga koristi M identičnih kopija sustava kao M uzorkovanih kanonskih ansambala na različitim temperaturama. Svaki iznos temperature jedinstven je i pridružen jednoj od M replika (T_1, T_2, \dots, T_M). Nad svakom od M replika neovisno se provodi jednostavna MC Boltzmann (p, T) pretraga uz odgovarajuću temperaturu.

U svom REMC algoritmu, Wang i suradnici uveli su dodatan korak, svaka replika p_i perturbira se, a pridružena vrijednost energije $E(p_i)$ arhivira u ansamblima P' i E' . Elitne replike u arhivima odabiru se primjenom *select_REMC_Replicas*(E', P') procedure. **Algoritam 2** predstavlja pseudokod za selekciju replika iz arhiva u implementaciji REMC metode uzorkovanja. Kao replike za zamjenu odabiru se posljednje “*numR*” konformacije koje su poslone u red za arhive, kao što je prikazano na **slici 1** (gore desno), gdje su “*numR*” uzorkovane strukture iskorištene kao replike za zamjenu (crvene točke).

Algoritam 2: SELECT_REMC_REPLICAS(E', P')

Input: E' – iznos energije u arhivima, P' – konformacijski ansambl u arhivima

Output: pe' – ansambl konformacije proteina odabrane elite

1 $i \leftarrow 0$

2 **while** $i < numR$ **do**

3 $pe' \leftarrow P'_{|E'|-i}$

4 $i \leftarrow i + 1$

5 **end**

6 **return** pe'

Trenutno stanje „ $numR$ “ replika odabranih iz arhiva možemo predstaviti kao ansambl konformacija proteina $pe' := (pe'_1, \dots, pe'_{numR})$, gdje je pe'_j konformacija replike j na temperaturi T_j . Tijekom zamjene replika, zamjenjuju se temperaturne vrijednosti susjednih replika s vjerojatnošću proporcionalnom njihovim energijama i razlici u temperaturi. Vjerojatnost prijelaza za neku trenutnu konformaciju pe'_i u perturbiranu konformaciju pe''_i (probni korak $Mover()$) određuje se na temelju Metropolis kriterija, (opisano u poglavlju o MC uzorkovanju, pogl. 3).

Kao što je već rečeno, zamjene se provide između dviju susjednih temperatura, T_i i T_j . Vjerojatnost zamjene ovisi o vrijednostima energija, $E(pe'_i)$ i $E(pe'_j)$, i inverznim temperaturama, β_i i β_j . Zamjena temperatura, odn. ponovno indeksiranje replika utječe na stanje ansambla replika pe' . Može se stoga općenitije definirati zamjena dviju replika i i j kao prijelaz iz trenutnog ansambla stanja pe' u izmijenjeno stanje pe'' . Definiramo trenutnu oznaku ili broj replike $l(pe'_i) = i$ za sve pe'_i . Vjerojatnost prijelaza iz trenutnog stanja ansambla pe' u izmijenjeno stanje pe'' zamjenom replika i i j definirano je kao:

$$P[pe' \rightarrow pe''] := P[l(pe'_i) \leftrightarrow l(pe'_j)] := \begin{cases} 1, & \Delta \leq 0, \\ e^{-\Delta}, & \text{u suprotnom.} \end{cases} \quad (6)$$

Vrijednost Δ umnožak je razlike u energiji i razlike u inverznoj temperaturi:

$$\Delta := (\beta_j - \beta_i) (E(pe'_i) - E(pe'_j)), \quad (7)$$

gdje je $\beta_i = 1/T_i$ inverzna temperatura replike i . Zamjene replika moguće su jedino između replika na susjednim temperaturama jer vjerojatnost prihvaćanja zamjene eksponencijalno pada kako se temperaturna razlika među replikama povećava.

Pseudokod za **algoritam 3** prikazuje detalje opisanog REMC postupka pretrage „ $numR$ “ replika za unaprijed definiran temperaturni raspon od $minT$ do $maxT$. Petlja “*while* $i + 1 < numR$ *do*” odnosi se na par replika koje se zamjenjuju i ukazuje na to da potencijalne zamjene uključuju parove (0, 1), (2, 3), (4, 5), itd., ali nikad parove (1, 2), (3, 4), (4, 6), itd., zbog čega nije zadovoljen „detaljan uvjet ravnoteže“ (engl. „*detailed balance condition*“) (vjerojatnosti prijelaza $i \rightarrow j \neq j \rightarrow i$). Nadalje, iz uvjeta za Δ očito je da se zamjena odbacuje ako je Δ veći od nekog praga (često 75, ali općenito ovisi o računalnoj arhitekturi) jer $e^{-\Delta}$ ne može nikada biti veći od bilo kojeg nasumičnog broja $mc_RG_uniform[0, 1]$. Iako se zamjena odbacuje, generirani nasumični broj se zadržava što čini algoritam kompjutacijski efikasnijim.

Algoritam 3: REMC($numR, numC, repackNth, minT, maxT$)

Input: p_0 – ansambl početnih konformacija, $numR$ – broj konformacijskih replika, $numC$ – broj koraka ciklusa, $repackNth$ – presloži bočni lanac nadopune sučelja receptora (engl. *receptor side-chain of interface padding*) svakih N koraka ciklusa, $minT$ – minimalna temperatura, $maxT$ – maksimalna temperatura

Output: p' – ansambl modificiranih stanja perturbiranih konformacija

```

1  $E' \leftarrow 0; P' \leftarrow 0$ 
2  $TStep \leftarrow (maxT - minT) / numR$ 
3 foreach temperature  $i$  in  $numR$  do
4   |  $T_i \leftarrow minT + TStep$ 
5 end
6 foreach cycle  $k$  in  $numC$  do
7   | foreach replica  $i$  in  $numR$  do
8     |  $p_i \leftarrow p_0$ 
9     | if  $i \% repackNth = 1$  then
10    | |  $p'_i \leftarrow packRotmersMover(p_i)$ 
11    | else
12    | |  $p'_i \leftarrow rotamerTrialsMover(p_i)$ 
13    | end
14    |  $MCBoltzmann(p'_i, T_i)$ 
15    |  $E' \leftarrow E(p'_i)$ 
16    |  $P' \leftarrow p'_i$ 
17  | end
18   $pe' \leftarrow select\_REMC\_Replicas(E', P')$ 
19   $i \leftarrow 0; j \leftarrow 0$ 
20  while  $i + 1 < numR$  do
21  |  $j \leftarrow i + 1$ 
22  |  $\Delta \leftarrow (\beta_j - \beta_i)(E(pe'_i) - E(pe'_j))$ 
23  | if  $\Delta \leq 0$  then
24  | |  $swapLables(pe'_i, pe'_j)$ 
25  | else
26  | |  $remc\_RG\_uniform \leftarrow U(0,1)$ 
27  | | if  $remc\_RG\_uniform \leq e^{-\Delta}$  then
28  | | |  $swapLables(pe'_i, pe'_j)$ 
29  | | end
30  | end
31  |  $i \leftarrow i + 2$ 
32  end
33   $p_0 \leftarrow 0; p_0 \leftarrow pe'$ 
34 end

```

§ 5. MONTE CARLO METODE UZ VIŠEKRITERIJSKU OPTIMIZACIJU ZAMJENE REPLIKA

Wang i suradnici razvili su novu inačicu REMC metode, Monte Carlo metodu uz višekriterijsku optimizaciju zamjene replika (engl. *multi-objective optimization replica exchange Monte Carlo*, MO-REMC), korištenjem više nedominiranih rješenja kao replika za zamjenu tijekom optimizacije na različitim temperaturama i na taj su način poboljšali konvergenciju REMC algoritma za uzorkovanje koja je povezana s odabirom replika za zamjenu. Njihove metode za odabir replika temelje se na informaciji o energiji iz Pareto fronte što poboljšava pretragu prostora stanja². Objašnjenja pojmova višekriterijskog optimizacijskog problema i Pareto optimalnosti dana su u poglavlju 5.1.

U MO-REMC metodi, za razliku od tradicionalnog REMC algoritma koji uglavnom uzorkuje kanonski ansambl stanja, uvedena je ovisnost o prošlosti korištenjem informacije višekriterijske optimalne Pareto fronte. Na taj je način olakšan odabir replika za zamjenu, koje čini set replika sličnih stanjima niže energije ali se istovremeno međusobno maksimalno razlikuju. Korištenje generirane Pareto fronte kao predloška reprezentativnih konformacija može pospješiti konvergenciju algoritma za uzorkovanje dostupnog konformacijskog prostora uključujući moguće strukture bliske nativnoj.

Cilj MO-REMC metode uzorkovanja povećanje je brzine konvergencije za dostupan konformacijski prostor. Spomenuta metoda koristi popis Pareto fronte ovisan o prošlosti kako bi se zadržao ograničen broj nedominiranih konformacija pronađenih REMC metodom. Svaka jedinka u arhivi generirana REMC metodom procjenjuje se pomoću binarnih kriterija: koraka u MC pretrazi i *TScore* vrijednosti perturbiranih konformacija u RosettaLigand programu. Kriterij MC koraka daje vremenske serije za proces pretrage, a kriterij *TScore*-ova informacijsku mapu dostupnog konformacijskog prostora ovisnu o prošlosti.

Rezultati eksperimenata pokazuju da elitne replike generirane pomoću višekriterijske Pareto fronte ovisne o prošlosti mogu povećati brzinu konvergencije za dostupni konformacijski prostor.

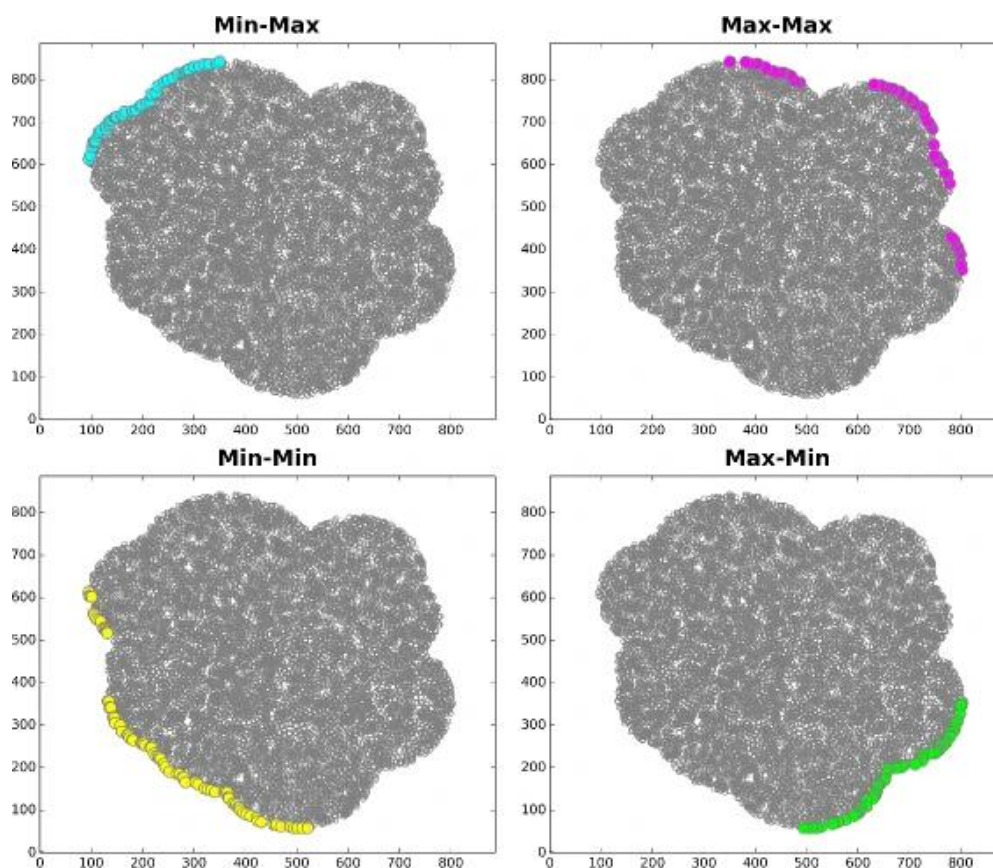
5.1. Višekriterijska optimizacija²

Višekriterijski optimizacijski problem (engl. *multiobjective optimization problem*, MOP) sastoji se od nekoliko kontradiktornih kriterija koji se moraju istovremeno optimizirati. Općenito, MOP se matematički definira kao:

Definicija 1. (Općeniti MOP): MOP minimizira $F(\vec{x}) = (f_1(\vec{x}), \dots, f_k(\vec{x}))$ s obzirom na $g_i(\vec{x}) \leq 0, i = 1, \dots, m, \vec{x} \in \Omega$. MOP rješenje minimizira komponentne vektorske funkcije $F(\vec{x})$, gdje je x n -dimenzionalni vektor varijable odluke ($x = x_1, \dots, x_n$) iz nekog prostora Ω . Vektor \vec{x} minimizira svaku komponentu $F(\vec{x})$, ili barem jednu, i funkcije komponente vektorske funkcije $F(\vec{x})$ trebale bi se moći izračunati za svaki \vec{x} .

Kriteriji iz **definicije 1** međusobno se isključuju jer nijedna točka ne maksimizira sve kriterije istovremeno. Najbolji kompromisi među kriterijima mogu se ostvariti u okviru Pareto optimalnosti.

Višekriterijski optimizacijski pristup pronalazi Pareto optimalan skup kojeg čini skup međusobno nedominiranih rješenja. U prostoru definiranom kriterijima, skup nedominiranih rješenja leži na površini poznatoj kao Pareto fronta. Skupovi nedominiranih rješenja su oni u kojima je svako rješenje bolje od svih ostalih barem po jednom kriteriju. Usporedbom svih rješenja u Pareto optimalnom skupu, uvijek je moguće poboljšati jedan od kriterija kako bi se ukupno postigla bolja dobit, bez da neki drugi postane lošiji. Razmatraju li se optimizacijski problemi s dva kriterija moguće je svaki kriterij minimizirati ili maksimizirati. Pristup Pareto fronte nudi metodu za pronalaženje podskupa obećavajućih rješenja utemeljenu na kriterijima. Na **slici 2** prikazan je primjer Pareto fronte s dva kriterija (obojene točke), gdje postoji kompromis između minimiziranja i maksimiziranja Pareto optimalnih točaka obiju, x i y , koordinatnih vrijednosti u *min-max*, *max-max*, *min-min* i *max-min* slučajevima. Grafovi prikazuju Pareto optimalan skup s diskretnim točkama za četiri različita slučaja i dva kriterija. Vidimo da Pareto optimalni skup uvijek sadrži rješenja s određenog ruba prostora pretrage dostupnog za diskretne točke.



Slika 2. Pareto optimalna rješenja iskorištena za pretragu četiri moguće kombinacije dvaju kriterija s diskretnim točkama. (Preuzeto iz ref. 2.)

5.2. Implementacija MO-REMC u RosettaLigand²

Algoritam 4 predstavlja pseudokod za računanje binarnih kriterija na temelju Pareto fronte arhiva u implementaciji MO-REMC metode uzorkovanja u RosettaLigand program. Svaki kriterij može biti minimiziran ili maksimiziran u skladu s vrijednostima *Boolean* varijabli $maxX$ i $maxY$. U ovom postupku, u prvom koraku (linije 1-6), sva rješenja x_0, \dots, x_{n-1} u arhivama predstavljaju alternative poredane prema povećanju odn. smanjenju kriterija X , koji može biti minimiziran ili maksimiziran. Neka je $pf' := \{x_0, y_0\}$ i $i := 1$, gdje $\{x_0, y_0\}$ označava kombinaciju koja sadrži prvu nedominiranu frontu. U sljedećem koraku (linije 8-17), za svaku kombinaciju u arhivama $\{x_i, y_i\} \in \{X, Y\}$, neka je $pf' := pf' \cup \{x_i, y_i\}$. Ako i $\{x_i, y_i\}$ nije dominiran nijednom kombinacijom s obzirom na kriterij Y koji je već minimiziran ili maksimiziran u pf' , onda se $\{x_i, y_i\}$ dodaje u pf' . U trećem koraku (linije 7-18), ponavlja se zapravo drugi korak dok ne ponestane kombinacija koje se mogu dodati u pf' . Zadnji korak predstavlja kraj iteracije, kada je $i=N$, gdje je N broj kombinacija u arhivama.

Algoritam 4: PARETOFRONTIER($X, Y, \max X, \max Y$)

Input: X – kriterij X , T – kriterij Y , $\max X$ – Boolean vrijednost maksimiziranog kriterija X , $\max Y$ – Boolean vrijednost maksimiziranog kriterija Y

Output: pf' – konformacijski ansambl Pareto optimalnih rješenja

```

1 if  $\max X = 1$  then
2   |  $inverse\_sorted(\{X, Y\})$ 
3 else
4   |  $sorted(\{X, Y\})$ 
5 end
6  $pf' \leftarrow \{x_0, y_0\}; i \leftarrow 1$ 
7 foreach  $\{x_i, y_i\}$  in  $\{X, Y\}$  do
8   |  $\{pair_{x-previous}, pair_{y-previous}\} \leftarrow pf'_{|pf'|-1}$ 
9   | if  $\max Y = 1$  then
10  |   | if  $pair_{y-previous} \geq y_i$  then
11  |   |   |  $pf' \leftarrow \{x_i, y_i\}$ 
12  |   |   end
13  |   else
14  |   | if  $pair_{y-previous} \leq y_i$  then
15  |   |   |  $pf' \leftarrow \{x_i, y_i\}$ 
16  |   |   end
17  |   end
18 end
19 return  $pf'$ 

```

Dodatno, u svakoj iteraciji MO-REMC metode uzorkovanja ne dobiva se samo zadnji set konformacija već novi set koristeći proceduru *select_MO – REMC_Replicas*(E', P'), za razliku od REMC metode, gdje se koristi *select_REMC_Replicas*(E', P'). Funkcija *select_MO – REMC_Replicas* omogućava odabir konformacija iz arhiva i zadnjeg seta „*numR*“ *min-min* scenarija Pareto optimalnih rješenja koji su nedominirani s obzirom na druge konformacije. Grafički su ova rješenja prikazana na **slici 1** (dolje lijevo), gdje se u posljednjem krugu posljednje „*numR*“ uzorkovane strukture koriste kao replike za zamjenu u MO-REMC metodi (crvene točke), a set *min-min* scenarija Pareto optimalnih rješenja označen je žutim točkama (dio točkaka je prekriven crvenim točkama). Spomenuti *min-min* scenariji Pareto optimalnih rješenja iz arhiva nude jednostavan i brz način konvergencije, što se koristi za dobivanje alternativnih setova za usporedbu iz arhiva. Pseudokod u **algoritmu 5** opisuje

postupak za odlučivanje treba li prihvatiti ili odbiti Pareto frontu, kao i za odlučivanje treba li odabrati replike za zamjenu ili ne.

Algoritam 5: SELECT_MO-REMC_REPLICAS(E', P')

Input: E' – iznos energije u arhivima, P' – konformacijski ansambl u arhivima

Output: pe' – ansambl konformacije posljednjeg odabranog „ $numR$ “

min-min scenarija Pareto optimalnih rješenja

1 $PF \leftarrow \text{paretoFrontier}(E'_{id}, E', \text{false}, \text{false})$

2 $i \leftarrow 0$

3 **while** $i < numR$ **do**

4 $pe' \leftarrow PF_{|PF|-i}$

5 $i \leftarrow i + 1$

6 **end**

7 **return** pe'

5.3. Implementacija hibridne MO-REMC u RosettaLigand²

U usporedbi s MO-REMC metodom, posebnost hibridne MO-REMC metode uzorkovanja (HMO-REMC) je samo u postupku odabira elitnih rješenja. U HMO-REMC metodi replike za zamjenu odabiru se iz rješenja u arhivima pri čemu se razmatraju nedominirana rješenja u kojima je i kriterij MC koraka i $Tscore$ minimiziran, kao i rješenja s maksimiziranim kriterijem MC koraka i minimiziranim kriterijem $Tscore$ vrijednosti. Na **slici 1** (dolje desno) može se vidjeti da se kao replike za zamjenu (crvene točke) u HMO-REMC metodi koriste nedominirana rješenja niže energije iz *min-min* i *min-max* scenarija Pareto optimalnih rješenja. *Min-min* scenarij Pareto optimalnih rješenja označen je žutim točkama, a *min-max* scenarij zelenim.

U svakom koraku odabira replika za zamjenu prednost imaju nedominirana rješenja niže energije i u *min-min* i *max-min* scenariju. Ako je broj rješenja manji od $numR$, što je broj replika iskorištenih za zamjenu, skup nedominiranih rješenja hibridizira se – i *min-min* i *max-min* scenariji nedominiranih rješenja koriste se iterativno za nadopunu skupa replika prema redoslijedu $Tscore$ vrijednosti.

Pseudokod za odabir HMO-REMC replika dan je u **algoritmu 6**.

Algoritam 6: SELECT_HMO-REMC_REPLICAS(E', P')

Input: E' – iznos energije u arhivima, P' – konformacijski ansambl u arhivima

Output: pe' – konformacijski ansambl odabranih elitnih replika

1 $PF_{ff} \leftarrow \text{paretoFrontier}(E'_{id}, E', \text{false}, \text{false})$

2 $PF_{tf} \leftarrow \text{paretoFrontier}(E'_{id}, E', \text{true}, \text{false})$

3 $i \leftarrow 0; j \leftarrow 0; k \leftarrow 0$

4 **while** ($i < |PF_{ff}| \ \&\& \ (j < |PF_{tf}|)$) **do**

5 **if** $E(PF_{ff(i)}) \leq E(PF_{tf(i)})$ **then**

6 $PF \leftarrow PF_{ff(i)}; i \leftarrow i + 1$

7 **else**

8 $PF \leftarrow PF_{tf(j)}; j \leftarrow j + 1$

9 **end**

10 **end**

11 **while** $i < |PF_{ff}|$ **do**

12 $PF \leftarrow PF_{ff(i)}; i \leftarrow i + 1$

13 **end**

14 **while** $j < |PF_{tf}|$ **do**

15 $PF \leftarrow PF_{tf(j)}; j \leftarrow j + 1$

16 **end**

17 **while** $k < \text{numR}$ **do**

18 $pe' \leftarrow PF_k; k \leftarrow k + 1$

19 **end**

20 **return** pe'

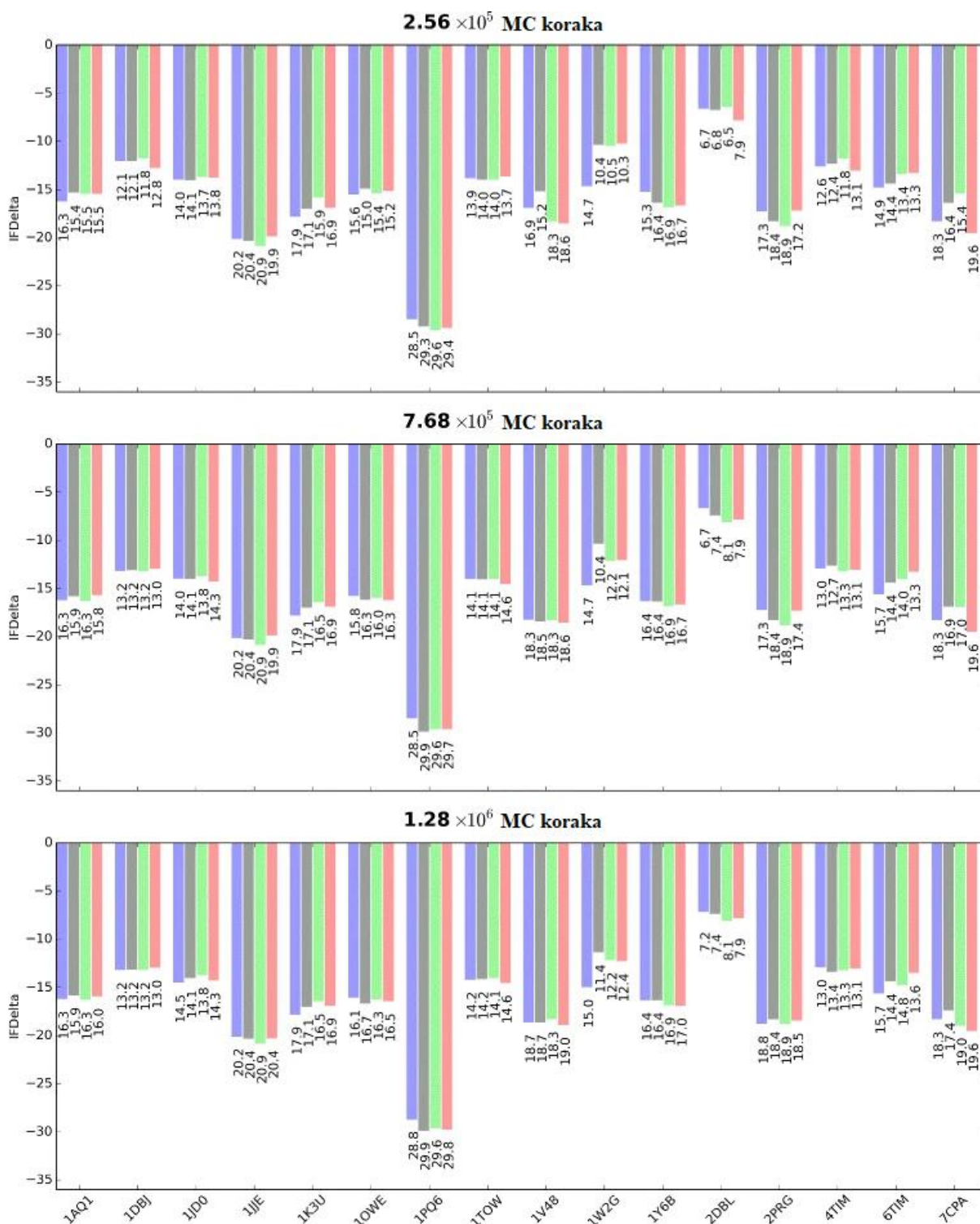
§ 6. USPOREDBA ODABRANIH MONTE CARLO METODA

Općenito se bolji rezultati dokinga ostvaruju uzorkovanjem konformacija nižih vrijednosti doking bodova (engl. *docking score*). Uvid u kvalitetu globalne izvedbe (engl. *global performance*) dokinga u RosettaLigand računalnom programu daje vezna energija sučelja proteina (engl. *binding energy interface delta, IFDelta*). Na **slici 3** vidimo da su sve odabrane metode (MC, REMC, MO-REMC i HMO-REMC) uspješno uzorkovale strukture slične nativnoj pri čemu su dobivene negativne *IFDelta* vrijednosti. Ipak, za isti broj MC koraka, unaprijedene metode uzorkovanja dale su rješenja s negativnijim vrijednostima *IFDelta* u usporedbi s klasičnom MC metodom. Nadalje, u duljim simulacijama, MO-REMC i HMO-REMC unaprijedene metode ostvaruju bolje rezultate uz negativniji konačni *IFDelta* u usporedbi s klasičnom MC i REMC metodom.

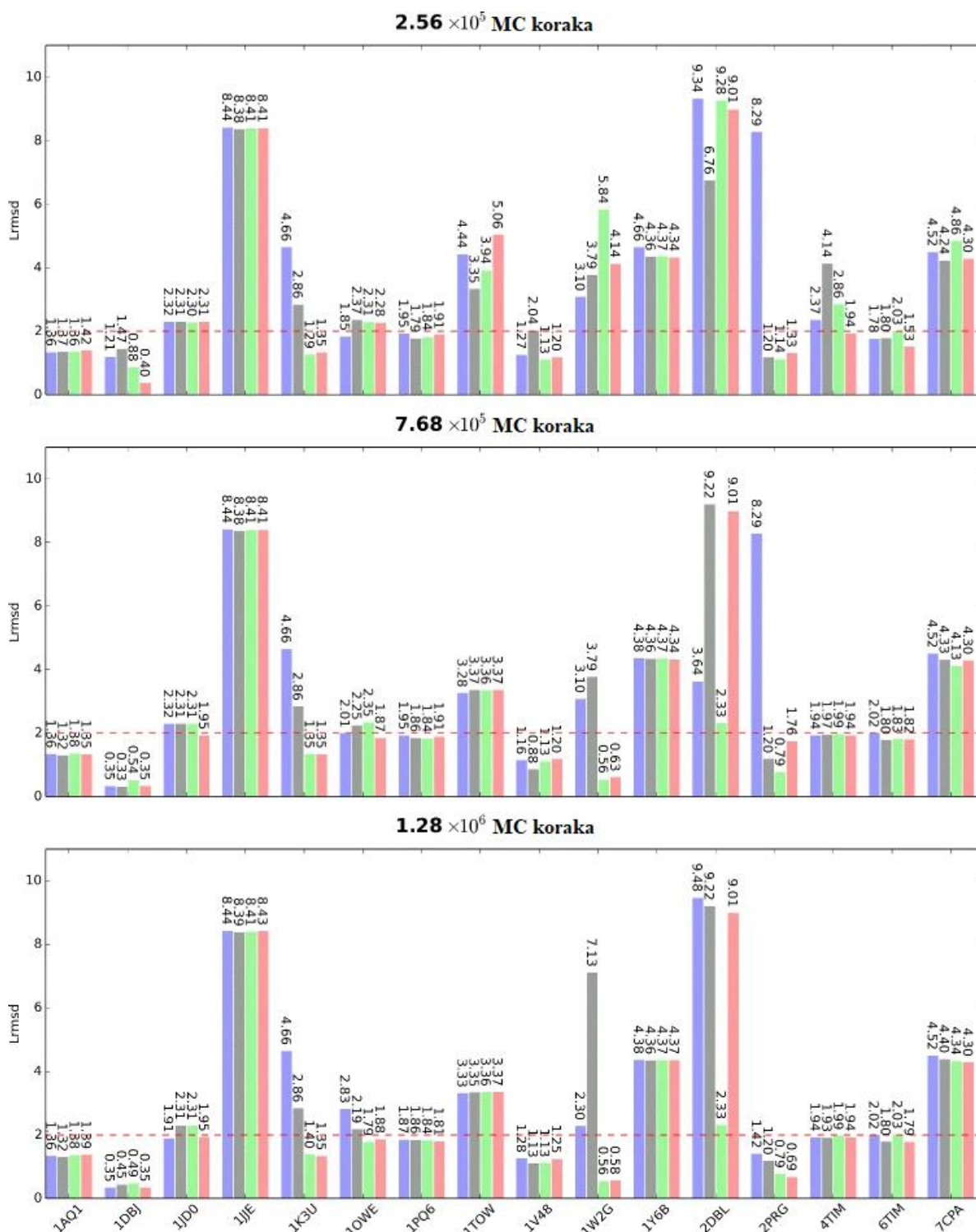
Sveukupna izvedba (engl. *overall performance*) dokinga može se opisati razlikom u RMSD liganda (engl. *ligand of RMSD, Lrmsd*). Pregled *Lrmsd* vrijednosti dobivenih za različite metode, receptore i duljine trajanja simulacije dan je na **slici 4**. Vidimo da u duljim simulacijama unaprijedene metode postižu bolje minimalne *Lrmsd* vrijednosti, gotovo savršene rezultate unutar 1 Å *Lrmsd*. Pri istom broju MC koraka, klasične MC metode ne mogu dati zadovoljavajuće uzorke složene proteinske fleksibilnosti.

U dovoljno dugim simulacijama promatrane ukupne vrijednosti energijske funkcije (engl. *energy function total score, TScore*) i *IFDelta* vrijednosti konvergiraju neovisno o primijenjenoj metodi (**slika 5**.) Na slici je prikazana i promjena *Lrmsd* vrijednosti, što omogućava provjeru utjecaja metode na rezultate. Procijenjene *Lrmsd* vrijednosti dobro se slažu s rezultatima *IFDelta* vrijednosti što potvrđuje da dobiveni rezultati ne ovise o odabranoj metodi. U svakom slučaju, unaprijedene metode uzorkovanja brže konvergiraju u usporedbi s klasičnom MC metodom.

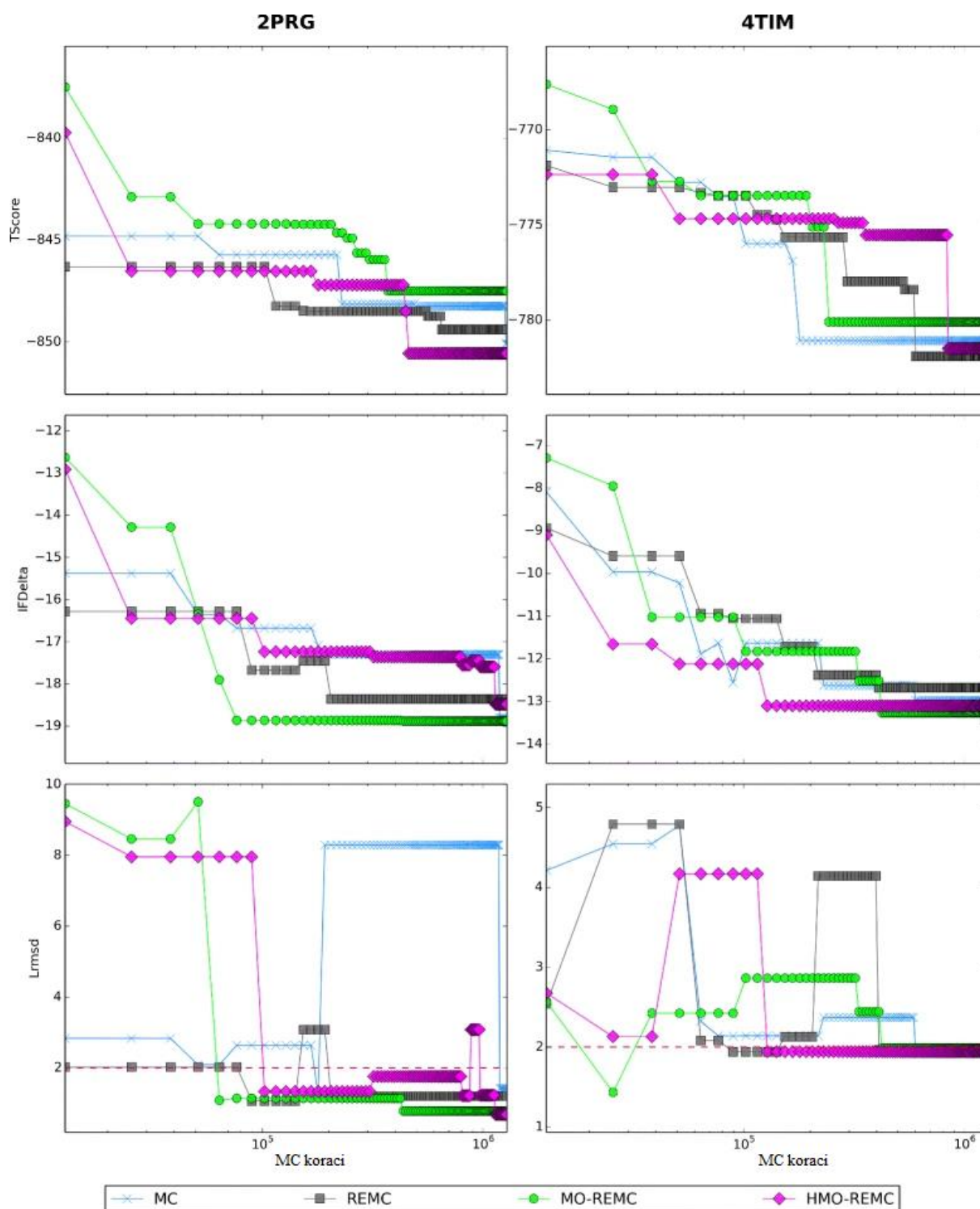
Valjanost razvijenih metoda (MO-REMC i HMO-REMC) provjerena je usporedbom više simulacija uz različite početne parametre. Rezultati prikazani na **slici 6** potvrđuju da ispitane metode mogu efikasno konvergirati u moguće strukture slične nativnoj.



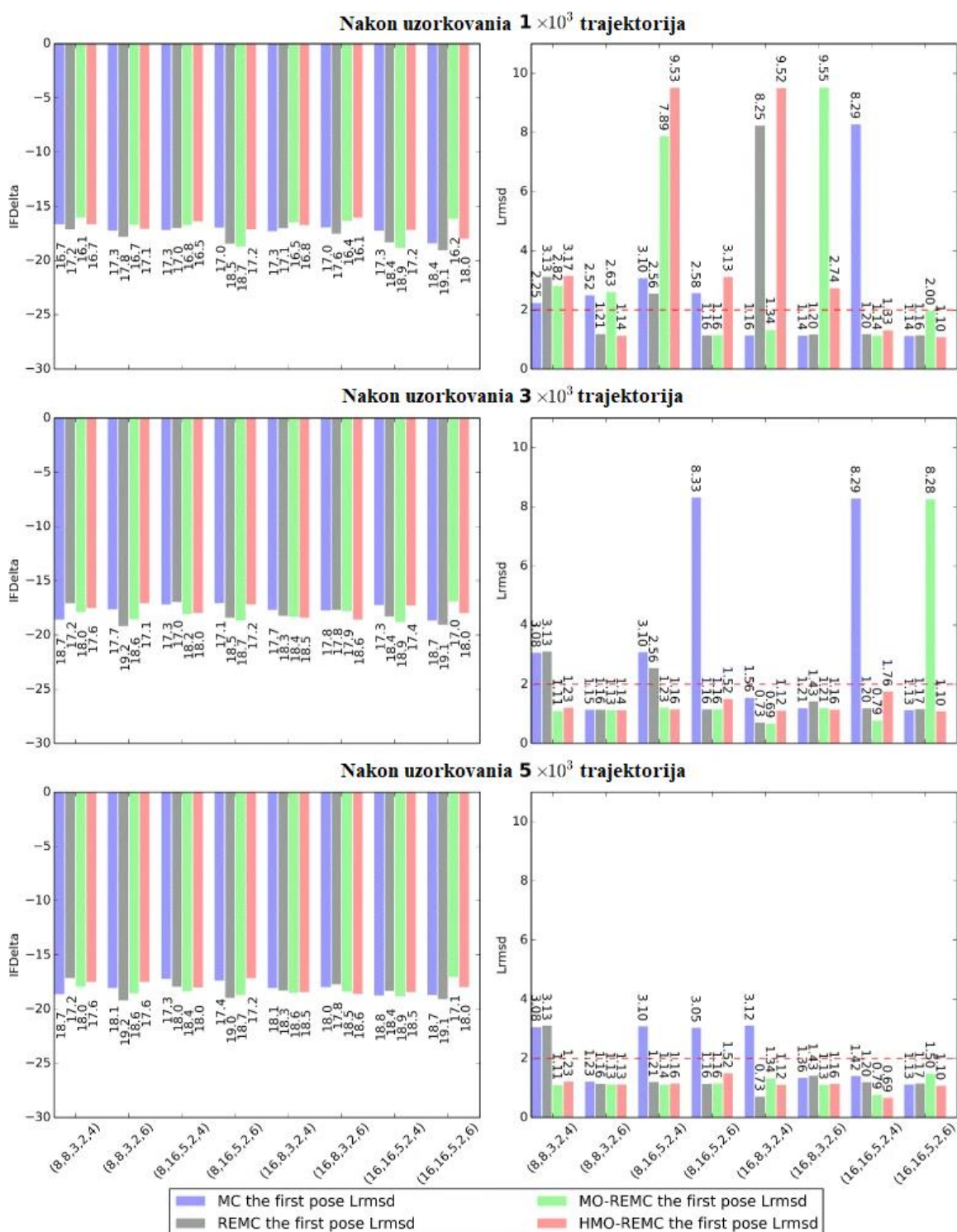
Slika 3. IFDelta vrijednosti dobivene MC (ljubičasto), REMC (sivo), MO-REMC (zeleno) i HMO-REMC (crveno) metodama uzorkovanja za 16 receptora (1AQ1, 1DBJ, 1JD0, 1IJE, 1K3U, 1OWE, 1PQ6, 1TOW, 1V48, 1W2G, 1Y6B, 2DBL, 2PRG, 4TIM, 6TIM, 7CPA) i 3 duljine trajanja simulacije ($2,26 \times 10^5$, $7,68 \times 10^5$ i $1,28 \times 10^6$ MC koraka; $MC\ koraci = numR \times numC \times$ broj trajektorija). (Preuzeto i prilagođeno iz ref. 2.)



Slika 4. *Lrmsd* vrijednosti dobivene MC (ljubičasto), REMC (sivo), MO-REMC (zeleno) i HMO-REMC (crveno) metodama uzorkovanja za 16 receptora (1AQ1, 1DBJ, 1JDO, 1IJE, 1K3U, 1OWE, 1PQ6, 1TOW, 1V48, 1W2G, 1Y6B, 2DBL, 2PRG, 4TIM, 6TIM, 7CPA) i 3 duljine trajanja simulacije ($2,26 \times 10^5$, $7,68 \times 10^5$ i $1,28 \times 10^6$ MC koraka; MC koraci = $numR \times numC \times$ broj trajektorija). (Preuzeto i prilagođeno iz ref. 2.)



Slika 5. Brzina konvergencije $Tscore$, $IFDelta$ i $Lrmsd$ vrijednosti u MC, REMC, MO-REMC i HMO-REMC metodama uzorkovanja za odabrane receptore (2PRG i 4TIM); $MC\ koraci = numR \times numC \times broj\ trajektorija$. (Preuzeto i prilagođeno iz ref. 2.)



Slika 6. Provjera valjanosti metoda. *IFDelta* i *Lrmsd* vrijednosti dobivene MC (ljubičasto), REMC (sivo), MO-REMC (zeleno) i HMO-REMC (crveno) metodama uz različite početne parametre ($(numR, numC, repackNth, minT, maxT) := npr. (8, 8, 3, 2, 4)$) za odabrani receptor (2PRG) i 3 duljine trajanja simulacije ($numR \times numC \times N$ MC koraka gdje je N broj trajektorija, $N = 1 \times 10^3, 3 \times 10^3$ i 5×10^3). (Preuzeto i prilagođeno iz ref. 2.)

§ 7. ZAKLJUČAK

Specifično vezanje liganda za receptor nužno je za pravilan rad metabolizma i provođenje signala u tijelu, a razumijevanje spomenute interakcije ključno je i za razvoj lijekova. U računalnom istraživanju interakcija proteina i malih molekula primjenjuju se tehnike molekularnog dokinga. Prvi korak uključuje generiranje svih mogućih konfiguracija, a zatim se one najznačajnije izdvajaju na temelju funkcija bodovanja.

Za dobivanje početnih konfiguracija često se koriste nasumične metode uzorkovanja struktura, između ostalog Monte Carlo algoritmi. U ovom radu opisane su jednostavne Monte Carlo metode (MC, algoritam 1), MC metode uz zamjenu replika (REMC, algoritmi 2 i 3), MC metode uz višekriterijsku optimizaciju zamjena replika (MO-REMC, algoritmi 3, 4 i 5) te hibridne MO-REMC metode (HMO-REMC, algoritmi 3, 4 i 6). Na temelju referentnih testova koje su proveli Wang i suradnici koristeći RosettaLigand protokol pokazalo se da primjena unaprijeđenih metoda može rezultirati boljim uzorkovanjem struktura prilikom molekularnog dokinga male molekule u aktivno mjesto proteina. Nadalje, MO-REMC i HMO-REMC metode brže konvergiraju prema rješenju u usporedbi s MC i REMC metodama uzorkovanja. Daljnjim unaprjeđenjem razvijenih metoda, primjerice nasumičnim odabirom replika iz arhiva ili uključivanjem Hamiltonijanske zamjene replika i sl., moglo bi se ostvariti još bolje uzorkovanje struktura.

§ 8. LITERATURNI IZVORI

1. C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd ed. (Wiley, Chichester, West Sussex, England, 2004).
2. H. Wang, H. Liu, L. Cai, C. Wang, and Q. Lv, *BMC Bioinform.* **18** (2017) 1–21.
3. J. Meiler, and D. Baker, *Proteins Struct. Funct. Genet.* (2006).
4. I. W. Davis, and D. Baker, *J. Mol. Biol.* **385** (2009) 381–392.
5. A. R. Leach, *Molecular Modelling: Principles and Applications*, 2nd ed. (Pearson Prentice Hall, Harlow, 2001).
6. D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, *Nat. Rev. Drug Discov.* **3** (2004) 935–949.
7. X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, *Curr. Comput. Aided-Drug Des.* **7** (2012) 146–157.
8. S. Y. Huang, and X. Zou, *Int. J. Mol. Sci.* **11** (2010) 3016–3034.
9. V. Salmaso, and S. Moro, *Front. Pharmacol.* **9** (2018) 1–16.
10. D. S. Goodsell, and A. J. Olson, *Proteins Struct. Funct. Bioinforma.* (1990).
11. G. M. Morris, D. S. Goodsell, R. Huey, and A. J. Olson, *J. Comput. Aided. Mol. Des.* (1996).
12. R. Abagyan, M. Totrov, and D. Kuznetsov, *J. Comput. Chem.* (1994).
13. C. McMartin, and R. S. Bohacek, *J. Comput. Aided. Mol. Des.* (1997).
14. M. Liu, and S. Wang, *J. Comput. Aided. Mol. Des.* (1999).
15. O. Trott, and A. J. Olson, *J. Comput. Chem.* (2010).
16. Y. Okamoto, *J. Mol. Graph. Model.* **22** (2004) 425–439.
17. L. Zheng, A. A. Alhossary, C. K. Kwoh, and Y. Mu, *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.* **1–3** (2018) 550–566.